SMOTE-FRST: A NEW RESAMPLING METHOD USING FUZZY ROUGH SET THEORY

E. RAMENTOL[1], N. VERBIEST[2], R. BELLO[3], Y. CABALLERO[1], C. CORNELIS[2,4] and F. HERRERA[4]

[1]*Department of Computer Science, University of Camagüey, Cuba,*
E-mail:enislayr@yahoo.es, yailec@yahoo.com
[2]*Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium,*
E-mail: Nele.Verbiest@UGent.be
[3]*Dept. of Computer Science, Universidad Central de Las Villas, Cuba,*
E-mail: rbellop@uclv.edu.cu,
[4]*Dept. of Computer Science and AI, University of Granada, Spain,*
E-mail: chriscornelis@ugr.es, herrera@decsai.ugr.es

In this paper, we introduce a new hybrid preprocessing method for editing imbalanced data. The algorithm we propose first resamples the training data using the Synthetic Minority Oversampling Technique (SMOTE) method, and subsequently applies an editing technique based on fuzzy rough set theory to the balanced training set. We evaluate the performance of our algorithm in an experimental study, using the C4.5 classifier as the learning algorithm. Statistical tests show the superiority of our method over state-of-the-art resampling methods.

Keywords: imbalanced data; resampling; classification; fuzzy rough sets

1. Introduction

An important data mining task is the problem of classifying imbalanced data.[?] These are datasets for which one of the classes is highly under- or overrepresented. In case of two classes, the problem we discuss in this paper, this means that there is one class with significantly more instances than the other class. The Imbalance Ratio (IR), defined by the number of instances in the majority class divided by the number of instances in the minority class, expresses to which extent a dataset is imbalanced: a dataset with IR equal to 1 is perfectly balanced, the higher the IR, the more imbalanced the dataset. Imbalanced datasets occur in many research fields, fraud detection, oil spill detection, text classification and medical applications are a few examples.

Standard classification methods are often biased towards the majority class, because they attempt to perform well w.r.t. global quantities such as classification accuracy, not taking the data distribution into account. As a result, examples from the majority class are generally correctly classified, whereas examples from the minority class are more often misclassified. In order to evaluate if a method is able to classify both instances from the minority and majority class well, the Receiver Operating Characteristic (ROC) curve that plots the rate of correctly classified minority instances against the rate of incorrectly classified minority instances can be used. The Area Under the ROC curve (AUC[2]) reflects the trade-of between correctly classified instances in the minority class and a high classification accuracy of instances in the majority class.

Several techniques have emerged to improve a classifier's performance under the data imbalance assumption. They can be divided into two classes depending on the level on which they operate, *viz.* learning algorithms level, and data level. In this paper, we work with the latter type: data level algorithms mainly focus on resampling instances, *i.e.* either generating synthetic instances from the minority class, removing examples from the majority class or both. Specifically, we focus on the Synthetic Minority Oversampling Technique (SMOTE[4]) methodology. As its name implies, this method generates synthetic instances, belonging to the minority class, to balance the training set.

In this paper, we attempt to improve SMOTE's performance by monitoring the quality of the generated synthetic instances. In an iterative process, after applying SMOTE, we remove synthetic minority instances, as well as original majority instances, that have a small membership degree to the fuzzy positive region. Such instances are considered to be noisy and hence are filtered out from the training data. The whole process (SMOTE instance generation + fuzzy-rough instance removal) is repeated until the training set is perfectly balanced. It is called SMOTE-FRST.

The remainder of this paper is structured as follows. In Section 2, we review basic preliminaries on fuzzy rough set theory, while in Section 3 we detail the methodology our proposal, the process SMOTE-FRST. In Section 4, an experimental study is set up to compare SMOTE-FRST to leading resampling techniques, using the well-known C4.5 classifier as the learning algorithm, and evaluating their performance on a large corpus of 44 datasets obtained from the KEEL dataset repository. Finally, in Section 5 we conclude.

2. Fuzzy Rough Set Theory

Rough set theory[6] provides a methodology for data analysis based on the approximation of concepts in a decision system $(X, \mathcal{A} \cup \{d\})$, in which X is a set of instances, $\mathcal{A}$ is a set of conditional attributes and d is the decision or class attribute. The theory revolves around the notion of (in)discernibility: the ability to distinguish between instances, based on their attribute values. When fuzzy rough sets are used, indiscernibility is typically modelled by means of a fuzzy tolerance relation R in X. In this paper, R is defined as, for x and y in X,

$$R(x,y) = T_L(\underbrace{R_a(x,y)}_{a \in \mathcal{A}}) \tag{1}$$

where T_L represents the Łukasiewicz t-norm given by $T_L(a,b) = \max(0, a+b-1)$ for $a, b \in [0,1]$, and $R_a(x,y)$ is given by

$$R_a(x,y) = \max\left(1 - \frac{|a(y) - a(x)|}{\sigma_a}, 0\right) \tag{2}$$

if a is a quantitative (real) attribute, and

$$R_a(x,y) = \begin{cases} 1 & \text{if } a(x) = a(y) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

when a is quantitative (discrete). Given R, the fuzzy-rough positive region of X is defined as the fuzzy set POS in X defined in this paper by

$$POS(x) = \min_{y \in X} I_L(R(x,y), R_d(x,y)) \tag{4}$$

where I_L is the Łukasiewicz implicator, given by $I_L(a,b) = \min(1, 1-a+b)$ for $a, b \in [0,1]$, and $R_d(x,y) = 1$ if x and y belong to the same class, and 0 otherwise. The idea of the fuzzy-rough positive region is that instances on the border of a class (i.e., for which there exists a similar instance in another class) will have a small membership value to POS compared to instances in the center of a class. This makes the fuzzy-rough positive region suitable to measure the quality of an instance as a typical representative of its class.

3. Methodology

In this section, we introduce SMOTE-FRST, our new hybrid editing method for imbalanced datasets. The algorithm consists of two stages, which are repeated until the training set is balanced, or until a maximum number of iterations is reached (10 in our experimental study):

(1) Apply SMOTE to introduce new synthetic minority class instances to the training set.
(2) Remove synthetic instances, or majority class instances, for which the membership to the fuzzy-rough positive region of the training set falls below a given threshold γ ($\gamma = 0.8$ in our experimental study).

Our hypothesis is that some of the introduced synthetic minority class instances may not be suitable for use in the learning phase, and hence should be eliminated. Similarly, majority class instances in the original training data that do not sufficiently belong to the fuzzy-rough positive region are removed. Note that we do not apply this procedure to the original minority class instances, since they are relatively sparse and are better left untouched.

4. Experimental Study

In this section we compare SMOTE-FRST to six well known preprocessing algorithms based on SMOTE:

- The SMOTE algorithm itself
- SMOTE with data cleaning using Tomek Links (SMOTE-TL[1])
- SMOTE with data cleaning using the Edited Nearest Neighbour technique (SMOTE-ENN[1])
- Borderline SMOTE, where only border instances are resampled (SMOTE-BL1[5])
- A variation on SMOTE-BL1 where the synthetic instances are closer to the minority class (SMOTE-BL2[5])
- SMOTE weighting the minority instances according to their safe-level (SMOTE-SL[3])

We apply these algorithms and the SMOTE-FRST algorithm to 44 datasets from the KEEL dataset repository[a], using the C4.5 classifier[7] as the learning algorithm. The datasets and their IR (we only use datasets with an IR higher than 9) are listed in Table 4.

For each experiment, we follow a 10 fold cross validation strategy: we divide the data in 10 folds and classify the instances of each fold using the remaining folds as training data.

In Figure 1, we show the average AUCs over all datasets for each method[b].

[a] http://sci2s.ugr.es/keel/datasets.php

[b] Due to space constraints, we cannot list all results in this paper, but they can be found at the url: http://users.ugent.be/∼nverbies/

Dataset	IR	Dataset	IR
yeast-2-vs-4	9.08	ecoli-0-3-4-vs-5	9
yeast-0-5-6-7-9-vs-4	9.35	ecoli-0-6-7-vs-3-5	9.09
vowel0	9.98	ecoli-0-2-3-4-vs-5	9.1
glass-0-1-6-vs-2	10.29	glass-0-1-5-vs-2	9.12
glass2	11.59	yeast-0-3-5-9-vs-7-8	9.12
shuttle-c0-vs-c4	13.87	yeast-0-2-5-7-9-vs-3-6-8	9.14
yeast-1-vs-7	14.3	yeast-0-2-5-6-vs-3-7-8-9	9.14
glass4	15.47	ecoli-0-4-6-vs-5	9.15
ecoli4	15.8	ecoli-0-1-vs-2-3-5	9.17
page-blocks-1-3-vs-4	15.86	ecoli-0-2-6-7-vs-3-5	9.18
abalone9-18	16.4	glass-0-4-vs-5	9.22
glass-0-1-6-vs-5	19.44	ecoli-0-3-4-6-vs-5	9.25
shuttle-c2-vs-c4	20.5	ecoli-0-3-4-7-vs-5-6	9.28
yeast-1-4-5-8-vs-7	22.1	ecoli-0-6-7-vs-5	10
glass5	22.78	ecoli-0-1-4-7-vs-2-3-5-6	10.59
yeast-2-vs-8	23.1	led7digit-0-2-4-5-6-7-8-9-vs-1	10.97
yeast4	28.1	glass-0-6-vs-5	11
yeast-1-2-8-9-vs-7	30.57	ecoli-0-1-vs-5	11
yeast5	32.73	glass-0-1-4-6-vs-2	11.06
ecoli-0-1-3-7-vs-2-6	39.14	ecoli-0-1-4-7-vs-5-6	12.28
yeast6	41.4	cleveland-0-vs-4	12.62
abalone19	129.44	ecoli-0-1-4-6-vs-5	13

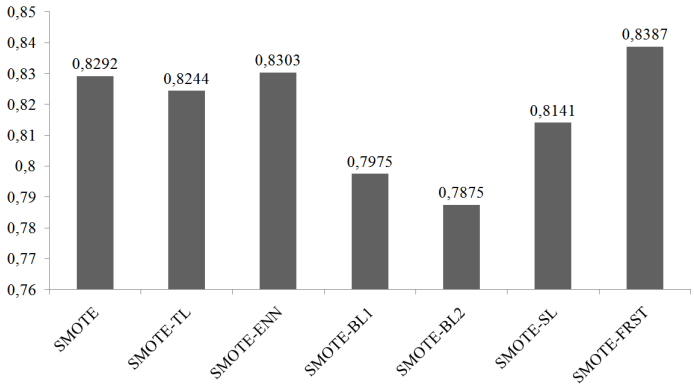

Fig. 1. Average AUC over all datasets.

SMOTE-FRST has the best average score. The improvement over SMOTE is small, so we must check if the differences are significant. We apply the Wilcoxon's signed ranks statistical test[8] to compare SMOTE-FRST against all other considered SMOTE methods. This is a non-parametric pairwise test that aims to detect significant differences between two sample means; that is, the behavior of the two implicated algorithms in the comparison. For each comparison we compute the sum of ranks of the Wilcoxons test in favor of SMOTE-FRST $R+$, the sum of ranks in favor of the other

methods $R-$ and also the p-value obtained for the comparison. The observed values of the statistics are listed in Table 2. The p-values are all lower than 0.10, which means that SMOTE-FRST outperforms the other SMOTE approaches at the 10 percent significance level.

SMOTE-FRST vs.	R+	R-	p-value
SMOTE	554	266	0.056
SMOTE-TL	667	236	0.007
SMOTE-ENN	572	289	0.066
SMOTE-BL1	833	113	0.00
SMOTE-BL2	805	98	0.00
SMOTE-SL	843	147	0.00

5. Conclusion

In this paper, we proposed a hybrid preprocessing technique for imbalanced datasets, the SMOTE-FRST process, which iteratively applies SMOTE to balance the dataset and uses fuzzy rough techniques to carry out editing on the synthetical and majority instances.
A preliminary experimental study shows that our proposed technique outperforms the existing SMOTE based techniques significantly.

References

1. G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29, 2004.
2. Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.
3. C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD09)*, volume 5476, pages 475–482, 2009.
4. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
5. Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *ICIC (1)*, pages 878–887, 2005.
6. Z. Pawlak. Rough sets. *International Journal of Computer and Information Science*, 11:341–356, 1982.
7. J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1993.
8. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, (6):80–83, 1945.